SIIM CAIMI26 Conference on Artificial Intelligence in Medical Imaging
OCT 26-27 | University of Pennsylvania, Philadelphia, PA

Large Language Model-Based Automated Detection and Turnaround Time Monitoring of Critical Results in a Pediatric Radiology Workflow

Pooja Pathak, MS, Lucile Packard Children's Hospital Stanford; Christina Vernazza; Linn Dee Doyle; Paul Fong; Zachery Seow, MBA, CRA, R.T.(R)(CT)(MR)(ARRT); Timothy Mak, MS; Richard Barth, MD; Ali Syed, MD; Shreyas Vasanawala, MD, PhD; Sergios Gatidis, MD

Introduction/Background
Critical results reporting and tracking are essential for quality assurance and workflow efficiency in radiology and are mandated by the Joint Commission. While commercial tools exist for adults, equivalent pediatric solutions are more challenging to develop, making in-house development particularly important. We aimed to develop and clinically deploy an LLM-based framework for automated detection and turnaround time monitoring of critical results in a pediatric radiology workflow.

Methods/Intervention
We defined six critical result categories: increased intracranial pressure, non-accidental trauma, malpositioned feeding tube, unsuspected pneumoperitoneum, portal venous gas, and tension pneumothorax. Finalized pediatric imaging reports and timestamps were exported to an on-prem database. Each report impression was processed by a locally deployed LLM (Llama 3.3, 70B, via Ollama) on NVIDIA H100 GPUs in a PHI-safe environment. A tailored prompt per category guided the model to return structured output with a yes/no classification and supporting evidence; an additional prompt identified documented evidence of communication. Turnaround time (TAT) was computed from exam end to final signature. We validated the framework against radiologist-adjudicated labels on 916 reports across six categories (5,496 report-category decisions), then deployed it retrospectively and prospectively, with results displayed in a Tableau dashboard.

Results/Outcome
The dataset comprised 182,047 finalized reports (122,962 XR, 29,097 US, 17,048 MRI, 6,186 CT, 5,433 FL, 1,321 NM). Mean patient age was 10.8 ± 7.4 years (51.7% male; n = 4,984). Validation yielded 351 true positives, 107 false positives, 1 false negative, and 5,037 true negatives, corresponding to sensitivity 99.7%, specificity 97.9%, and NPV 99.9%. Prospective implementation from August 2024 through July 2026 yielded 1,297 flagged critical results: 892 malpositioned feeding tubes, 158 portal venous gas, 109 tension pneumothorax, 89 increased intracranial pressure, 40 unsuspected pneumoperitoneum, and 9 non-accidental trauma. TAT and communication documentation were tracked continuously and remained consistent throughout deployment.

Conclusion
A locally deployed, PHI-safe LLM framework reliably automated critical results detection and TAT monitoring in pediatric radiology, achieving 99.7% sensitivity in validation and continuously monitoring 182,047 reports over 24 months of continuous clinical use.

Statement of Impact

Open-weight models on local infrastructure make continuous monitoring of every finalized report achievable in-house, without commercial tooling or external data transfer.

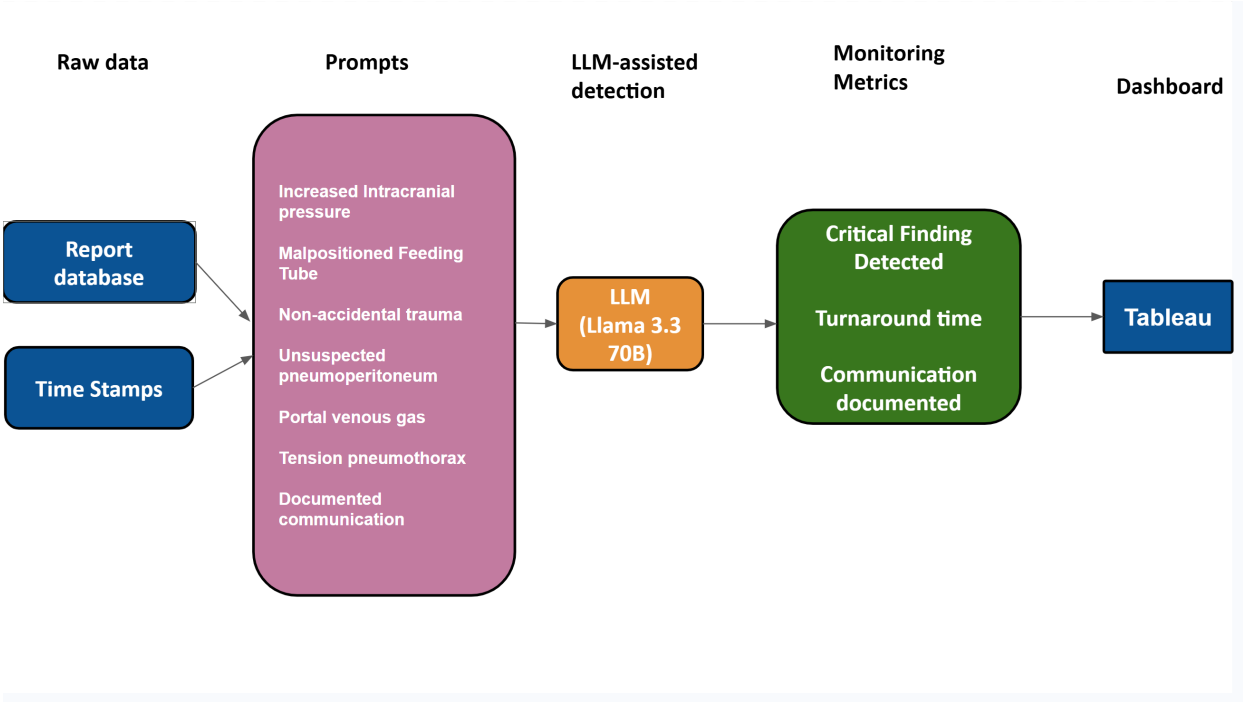

Fig 1: LLM-based pipeline for automated detection, turnaround time measurement, and communication tracking of critical results in pediatric radiology reports.

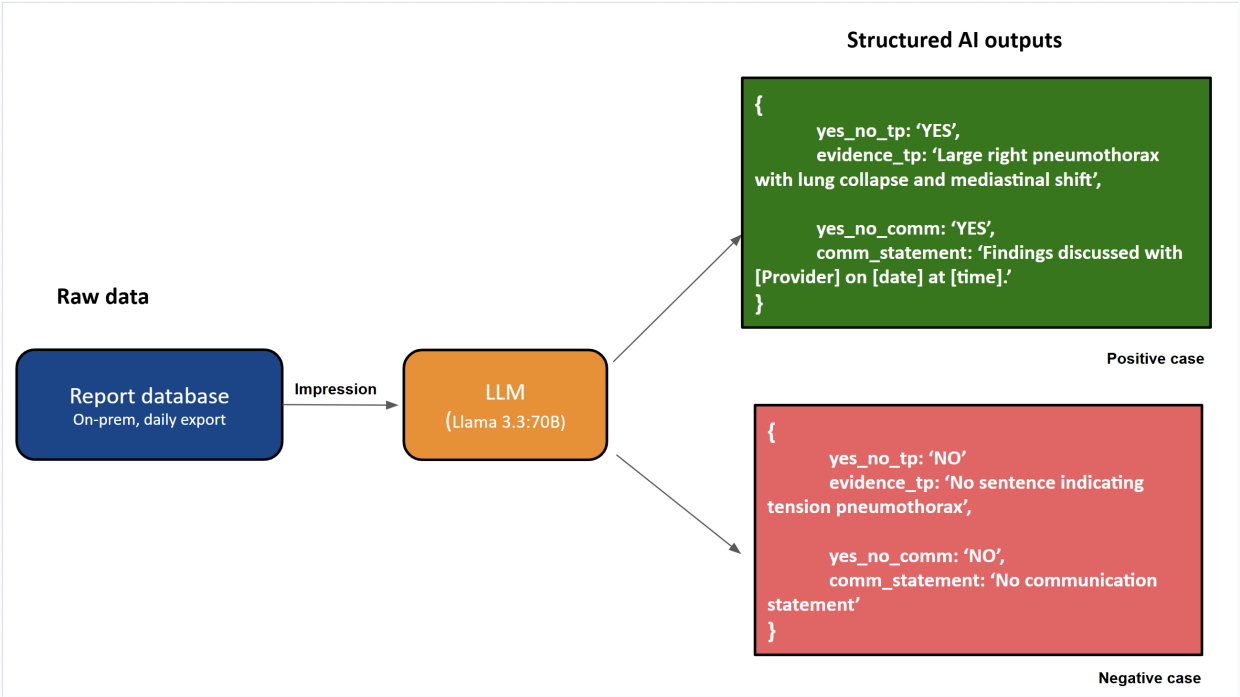

Fig 2: Structured LLM output for positive (top) and negative (bottom) tension pneumothorax cases, showing the binary classification, supporting evidence, and communication documentation returned per category.

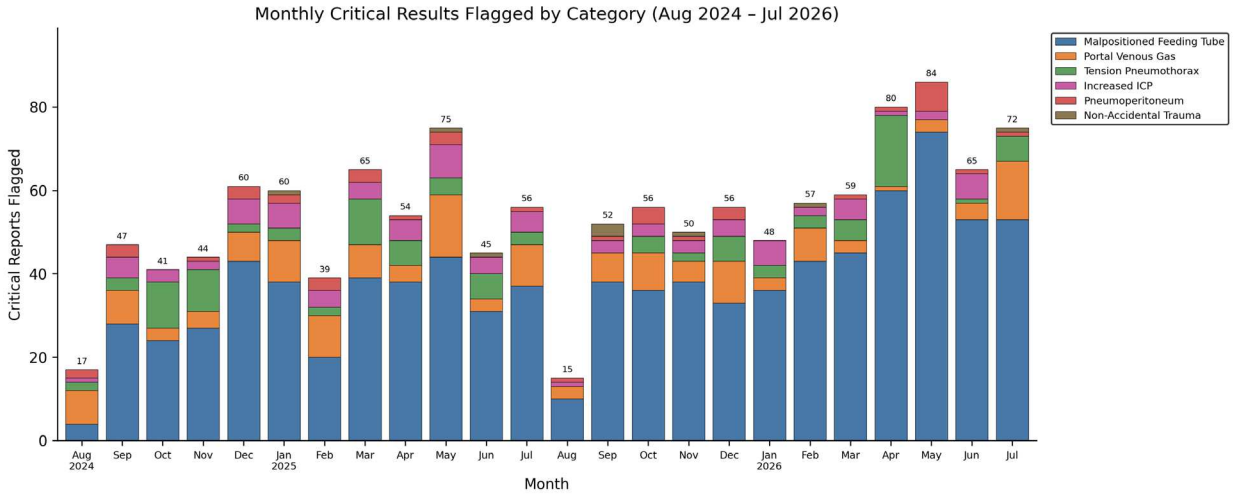

Fig 3: Monthly critical results flagged by category across the 24-month prospective deployment (August 2024 - July 2026; n = 1,297). Malpositioned feeding tube was the predominant critical result category throughout the deployment period.

Keywords

Large Language Models; Critical Results; Pediatric Radiology; Clinical Deployment; Natural Language Processing; Structured Data